

## ***Spitzer* Space Telescope Data Processing and Algorithmic Complexity**

Mehrdad Moshir

*Spitzer Science Center, California Institute of Technology, Pasadena,  
CA 91125, Email: mmm@ipac.caltech.edu*

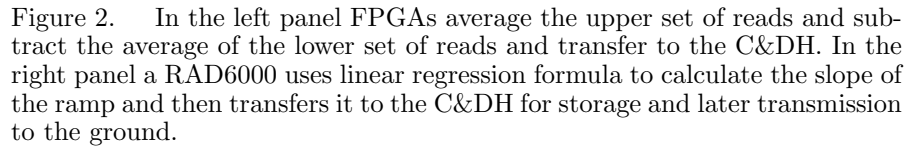
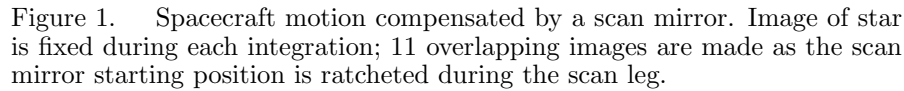
**Abstract.** Automated reduction of a very large body of data from the *Spitzer* Space Telescope requires an intricate and very flexible software system to support more than 50 different pipelines executing on a large distributed computing cluster. Additionally, in order to correct many instrumental signatures and calibration features, a variety of complex algorithms need to be utilized to process and distribute calibrated images and spectra from nearly 5 million instrument frames that are collected by the observatory every year. A sample of the complexities that needed to be accommodated both in system architecture and design as well as signatures encountered and associated algorithms will be described.

### **1. Overview**

The *Spitzer* Space Telescope, launched in August of 2003, has been operating very successfully for well over a year and has returned a significant body of new data in the infrared domain (Werner et al. 2004). The observatory is capable of generating close to 5 million distinct instrument frames every year, and of necessity such a large volume of data needs to be processed in a lights-out fashion. The automated processing of such a voluminous and varied number of datasets in a timely fashion imposes certain complexities on the design and implementation of the system infrastructure as well as its individual modular components. Data collection approach on-board the spacecraft injects further complexity into the ground software and pipelines. In Section 2 we will briefly discuss the data collection environment and its impact on pipeline design. In Section 3 we will describe the operational environment related to downlink processing. In Section 4 we will address the requirements on the system and the approaches adopted. Finally in Section 5 we will touch on some of the signatures and effects that need to be corrected by the ground software.

### **2. Data Collection**

The instruments on board *Spitzer* can collect data in the range of  $3.5\mu\text{m}$  to  $160\mu\text{m}$  with the spacecraft either pointed inertially or with the spacecraft performing a sky scan at a constant rate while an internal mirror compensates the motion and keeps the sky fixed on the focal plane. The latter mode of data collection permits large surveys with little pointing overhead, a cartoon of a typical scan observation by the MIPS instrument is seen in Figure 1. The three instrument suites consist of many different array technologies, InSb, Si:As, Si:Sb,



All instruments perform non-destructive reads of the respective arrays. However, due to the limited on-board storage and ground communication bandwidth it is not possible to send down all samples. Depending on the instruments (Ge instruments excluded for now), either FPGAs or a RAD6000 computer perform some “data compression” before transferring the data to the Command and Data Handling computer (another RAD6000). The limited computational power of the RAD6000 does not allow sophisticated or robust estimation, and some undesired side effects may result that will need correction in the pipelines, as will be seen later. An example of data compression by two of the instruments is seen in Figure 2. One of the drawbacks of on-board data compression is that the dispersion in the data is not retained. The issue of how uncertainties are dealt with in *Spitzer* has been discussed previously (Moshir et al. 2003). The data collected by the instruments are compressed and stored by the C&DH and sent to the ground on a pre-determined basis; typically 1GB of compressed data (better than 2X compression) are transferred to the ground every day.

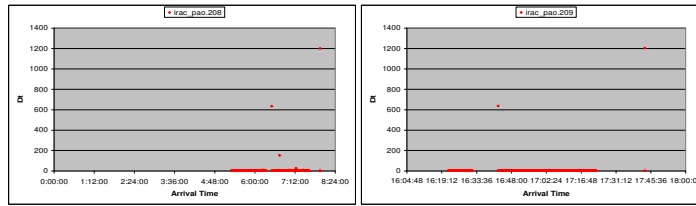


Figure 3. Each point refers to the time separation between one DCE arrival and the next. The left panel shows a morning DSN contact and the right panel shows a typical afternoon contact (both for IRAC).

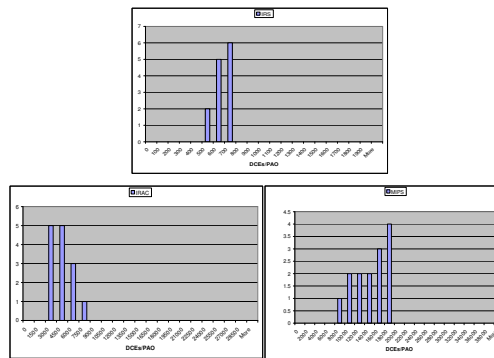


Figure 4. Average DCE rates per day vary by a factor of 20+. To date the peak DCE rates have been 1,160/day, 12,500/day and 35,700/day for IRS, IRAC and MIPS, respectively (1 day  $\sim$  2 PAOs).

### 3. Operating Environment

The Observatory is radiated a set of observations to be performed autonomously over a given time period; such a period is typically 12 hours long and is called a Period of Autonomous Operation (PAO). Only one instrument operates at a time during a PAO, several such consecutive PAOs form an instrument Campaign. The raw instrument data are referred to as Data Collection Events (DCEs). The DCEs are assembled from decompressed telemetry packets at JPL and sent to the SSC for processing by pipelines. Details of these concepts have been given previously (Moshir 2002).

While a station contact is about 30 minutes long, the data arrival at the SSC is stretched out due to resource sharing at JPL among many other space missions. Typical data arrival duration may range from 1 1/2 to 2 1/2 hours as in Figure 3; the average inter-DCE arrival time is close to 2 seconds (for the cases shown).

As the data arrive at the SSC the ingestion process takes place automatically (event-driven via data arrival). During a day anywhere from 1,100 DCEs to 35,000+ DCE are received. The distribution of data rates for the first 10 months of routine operations is seen in Figure 4. As of campaign 13 of IRS, IRAC and MIPS a total of 102,699, 1,059,065 & 3,005,579 DCEs had been received ( $\sim$  10 months of routine operations).

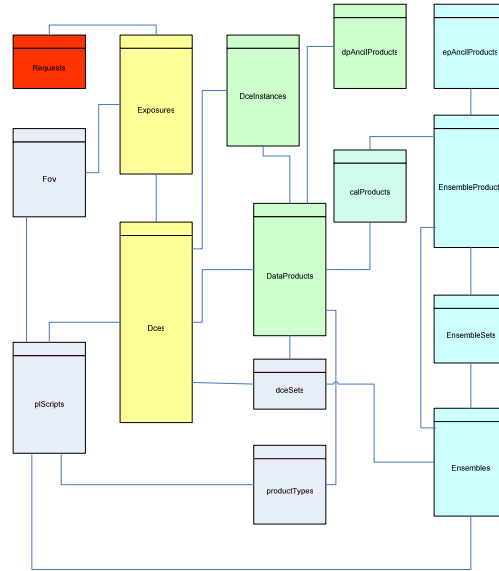


Figure 5. A subset of DB schema just for tracking and running of pipelines. For the whole set of 14 tables, containing 180+ columns, the average growth rate is in excess of 234,000 rows per day

#### 4. System Requirements

The downlink system needs to enable several fundamental capabilities, such as: 1) Book-keeping in a distributed asynchronous system. 2) Automated data reduction of large data sets. 3) Automated calibration retrieval. 4) Support for many pipelines (50+) for many data modes. 5) Rapid reduction of arriving data (5X real time). 6) Timely quality assessment of reduced data. 7) Rapid re-wiring of pipelines when needed (specially during the IOC and SV phases). 8) Re-use of automated pipeline modules in interactive tools. 9) Enhancement of spacecraft pointing reconstruction using science data.

For book-keeping of records and pipeline transactions, among other needs, the system relies on a complex relational DB that contains 1) Uplink/scheduling information and instrument settings. 2) Status of received data and pipeline information as well as pointers to data in the file system. 3) Calibration metadata and file pointers. 4) QA information for the pipelines.

As an illustration a subset of the database schema that deals *only* with tracking and running multiple pipelines is seen in Figure 5. The tables shown have 180+ columns. The main driver, the DCE instances table grows at an average rate of 12,000+ rows per day. For automation of the system it is required to route different data to different pipelines, this is accomplished by creation of “job manifests” that allow the pipeline executive to route the data. Pipelines that require an *ensemble* of data to perform tasks such as mosaicing, dark estimation, etc. require a way to associate related data for processing. This is accomplished by an automated ensemble maker (Laher & Rector 2004).

Given the job manifests and ensemble sets, the pipelines get started after the ingestion process has been invoked upon data arrival. Pipelines run on

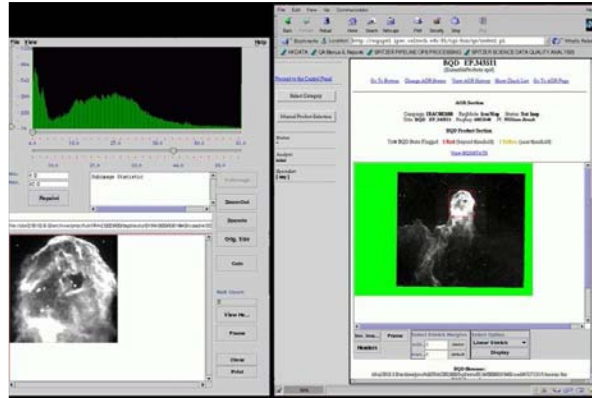


Figure 6. A web-based tool allows the DQA team to inspect any processed dataset to assess validity for release to the observers.

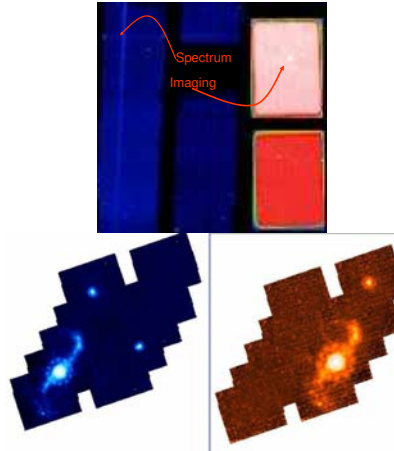


Figure 7. Re-use of IRS instrument for imaging; re-use of software to perform mosaicing of resulting data at  $16\mu\text{m}$  and  $22\mu\text{m}$ .

many “drones” for each data type and deposit the results in the “sandbox”. The pipelines are designed with flexibility in mind and are easy to re-configure (Brandenburg et al. 2004). One set of pipelines performs calibration reduction (darks, flats, etc.), these pipelines populate the Calibration Transfer tables in the database; the “CalTrans” system relies on this database for serving the proper calibration terms to the science pipelines, it ensures that science pipelines use calibration of the epoch in each case (Lee et al. 2004). To support QA functions all pipelines generate significant statistics for each processed DCE and deposit them into QA tables for each instrument. These tables allow the DQA team to easily access and certify data products for release to the observers using a web-based tool (Narron et al. 2003) as seen in Figure 6.

The flexibility of pipeline components allows rapid development of *new* pipelines. For example the IRS instrument, while designed for spectroscopy, also has two apertures in one array that can image the sky. Starting in Cycle 2 of *Spitzer*

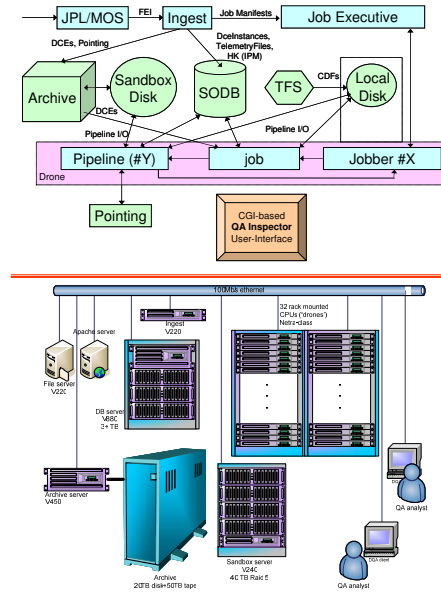


Figure 8. The pipeline data flow diagram (upper panel), and the hardware architecture to support it (lower panel); behind a firewall.

Call for Proposals, users can request imaging at  $16\mu m$  and  $22\mu m$ . To create the pipeline for reducing such data was very simple. Several modules that performed similar tasks for other instruments were re-wired to create a mosaicing capability for these observations, an example seen in Figure 7.

Early in the design of the pipelines it was anticipated that while automation of data reduction was applicable to almost all cases, there would be instances where the *intent* of the observer could not be determined via artificial intelligence and the same pipeline modules would need to support interactive usage. The domain of spectroscopy and spectrum extraction fall into such cases. The individual pipeline modules are easily pluggable into interactive tools to allow supervised spectral extractions by an observer (Hesselroth 2004).

To enable processing the data at least five times faster than data acquisition rate, using limited funds for computing machinery (“Netra class” blades), necessitated a large cluster of CPUs. These “drones” are configured to process any two pipelines concurrently, the pipelines get their data from the archive and process them locally then deposit the results into the sandbox for later archiving. The software and hardware architecture to allow this is seen in Figure 8.

## 5. Instrument Signatures, Data Collection Environment and Algorithmics

The instruments on-board *Spitzer* are all state-of-the-art and not flown before. Unlike the CCDs that have been very well characterized, there are many signatures that need to be understood, characterized and removed. To name a few of complications, we observe that there are on-board “processing” features; we

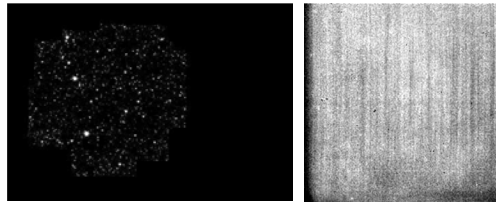


Figure 9. A low background region of sky as seen by IRAC (left). Carefully dithered observations of the same region and rejecting point sources then optimally combining the results yields a very good “sky dark” image (right).

have a shutter-less operation; there is optical cross-talk; instruments are non-linear; there are latent images; there are time-dependent readout signatures; there is readout droop; there is debris on one pick-off mirror that combined with scan mirror motion (Figure 1) results in non-stationary flat-fields; the Ge detectors have time-dependent, background-dependent responsivity changes needing a “self-calibrating” observation strategy; there is long-term memory & “action at a distance”; there are optical distortions; the spacecraft pointing, while good, needs to be further refined using science contents, and so forth.

An example of on-board processing features is that the RAD6000 computer used for partial data reduction (aka “compression”) does not have the capacity to account for ramp non-linearities, sample correlation (Fowler 2004) or ramp saturation or a particle hit causing a ramp discontinuity. Nevertheless the pipelines have been designed to correct for these features (Masci et al. 2004).

To operate imaging instruments without deploying the shutter and still obtaining “dark”-corrected calibrated products is non-trivial. To estimate the dark offset one may consider pointing the telescope at a low background part of the sky, say towards the North Ecliptic Pole; unfortunately there are background stars and galaxies everywhere easily visible to the sensitive eyes of the IRAC instruments! However, clever dither patterns and rejecting point sources and outliers and combining the results yields excellent “sky dark” images that, combined with “lab darks”, effectively act as though a shutter had been deployed (see Figure 9).

One other example of non-trivial data reduction is pointing refinement. There are some significant drivers for achieving maximal pointing knowledge. Good pointing allows source identification and future follow-on observations, it results in higher S/N in mosaics and point source detections, and it also facilitates potential super-resolution methods (Backus et al. 2004). In the process of pointing refinement it is necessary to remove distortions, since uncorrected distortions will limit how well pointing could be corrected. To characterize distortions a large body of data needs to be analyzed and deviations fitted to a multinomial in array coordinates; these coefficients appear in all of the *Spitzer* products (Shupe et al. 2004). The *Spitzer* imaging pipelines account for distortions (Makovoz & Khan 2004), and the resulting mosaics are distortion-free. An example of the type of distortions encountered and corrected is seen in Figure 10.

Once the distortions in each individual image have been corrected, the point sources within overlapping frames are matched to a good IR astrometric catalog, such as 2MASS, and then the pointing correction for each individual image

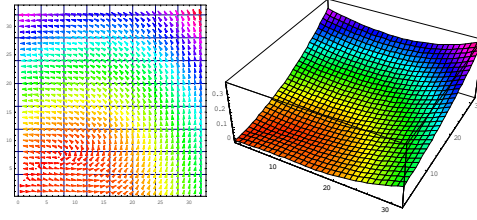


Figure 10. Optical distortion illustration for MIPS Ge, for *one* mirror position.

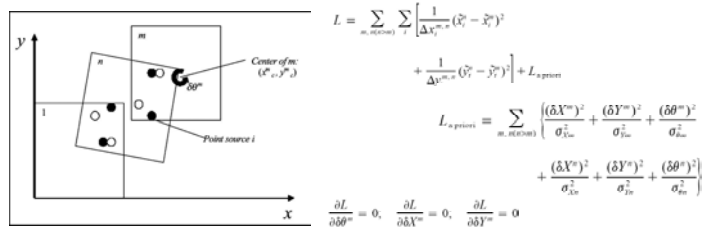


Figure 11. Distances of same source apparitions in overlapping images are minimized while using the prior knowledge about the expected pointing errors, by minimizing the cost function  $L$ .

is obtained by minimizing a cost function  $L$ , defined by the residual distances between apparitions of same source in overlapping frames as well as against astrometric counterparts while including prior knowledge of the expected pointing uncertainties, as shown in Figure 11.

When such minimization of the cost function takes place, typically it becomes necessary to invert matrices that are several thousand by several thousand; these matrices are usually sparse, and standard libraries are invoked for their solutions (Masci, Makovoz, & Moshir 2004). Pointing corrections for IRAC are significant and reduce the errors from just under  $1''$  to  $0.1'' - 0.2''$ ; pointing refinement manifests itself by increasing the S/N in a mosaic as well as sharpening it, Figure 12.

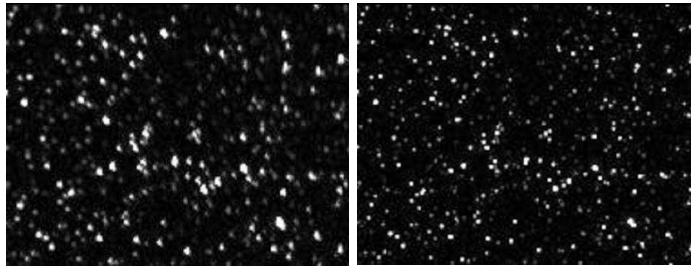


Figure 12. The mosaic on the left was made without pointing refinement; the one on the right was made from same data after performing pointing refinement.



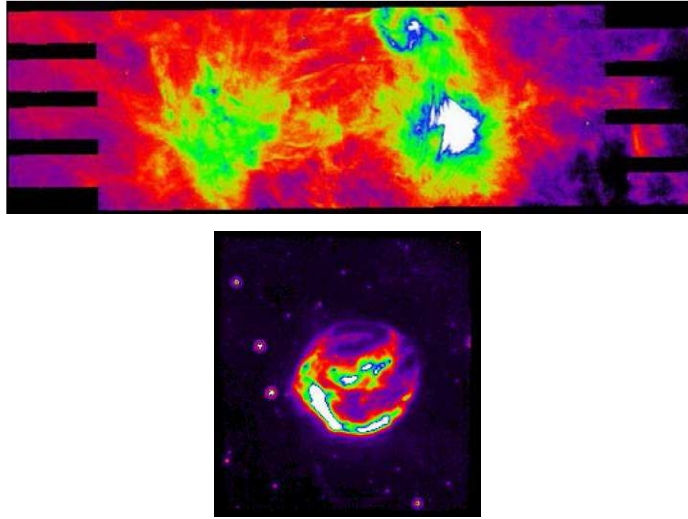


Figure 13. The top mosaic was made from a scan map observation (Perseus molecular cloud from the c2d Legacy Program). Lower mosaic was obtained in dithered mode (Kepler supernova remnant); both at  $24\mu m$ .

### 5.1. Scanning, Data Redundancy, and Self-Calibration

In Section 2 pointing modes of observing were discussed. The MIPS suite of instruments is able to observe both in dithered pointed or scan map modes. The complexities of accounting for mirror-dependent distortions and mirror-dependent flat fields are taken care of by the pipelines, and either large area or small area mosaics are obtained (Figure 13). One area that poses new challenges is the reduction of MIPS Ge data because of the well-known tendency of Ge detectors to have non-stationary responsivity. For this reason the Ge instruments provide a self-calibrating observation strategy. At regular time intervals a stimulator flash is turned on and the arrays are exposed to a highly repeatable photon flux. In Figure 1 the redundant coverage of the same point on the sky by different parts of the array was shown for a medium speed scan map. By using such redundancy and tracking the response to stimulator flashes, it is possible to mitigate the Ge responsivity variations, and the data are “self-calibrated” within each observation set (Henderson et al. 2004; Pesenson et al. 2004).

## 6. Summary

The approach to designing a distributed computing environment with very heavy reliance on a complex database is seen to be capable of meeting the challenges of reducing close to 5 million images per year from *Spitzer* as well as reprocessing them at the same time as the real time arriving data. Although the computing engine for the database exceeds the power and speed of all computing drones used for pipeline processing, the requirement of having a centralized database of all mission data has been met. The approach to modularized and generic design with an eye toward re-use of code in interactive tools has proved to be a

significant resource saving measure. The capability to rapidly refine the pipelines during the IOC and SV phases allowed the pipelines to react to on-orbit realities very quickly to produce the best calibrated products. As a result the pipeline products are currently being used for immediate scientific research soon after the start of routine operations (e.g., the issue of *ApJ* (Supp) 2004, 154, 1).

**Acknowledgments.** The program described here owes its success to the contributions of many, particularly the efforts of the Downlink development team and the understanding and support of Bill Green during his tenure as the SSC manager. This work was carried out at the *Spitzer* Science Center, with funding from NASA under contract 1407 to Caltech and the Jet Propulsion Lab.

## References

- Backus, C., et al. 2004, in ASP Conf. Ser., Vol. 347, ADASS XIV, ed. P. L. Shopbell, M. C. Britton, & R. Ebert (San Francisco: ASP), 61
- Brandenburg, H., et al. 2004, in ASP Conf. Ser., Vol. 347, ADASS XIV, ed. P. L. Shopbell, M. C. Britton, & R. Ebert (San Francisco: ASP), 575
- Fowler, J. 2004, in ASP Conf. Ser., Vol. 347, ADASS XIV, ed. P. L. Shopbell, M. C. Britton, & R. Ebert (San Francisco: ASP), 449
- Henderson, D., et al. 2004, in ASP Conf. Ser., Vol. 347, ADASS XIV, ed. P. L. Shopbell, M. C. Britton, & R. Ebert (San Francisco: ASP), 590
- Hesselroth, T. 2004, in ASP Conf. Ser., Vol. 347, ADASS XIV, ed. P. L. Shopbell, M. C. Britton, & R. Ebert (San Francisco: ASP), 296
- Laher, R., & Rector, J. 2004, in ASP Conf. Ser., Vol. 347, ADASS XIV, ed. P. L. Shopbell, M. C. Britton, & R. Ebert (San Francisco: ASP), 39
- Lee, W., et al. 2004, in ASP Conf. Ser., Vol. 347, ADASS XIV, ed. P. L. Shopbell, M. C. Britton, & R. Ebert (San Francisco: ASP), 594
- Makovoz, D., & Khan, I. 2004, in ASP Conf. Ser., Vol. 347, ADASS XIV, ed. P. L. Shopbell, M. C. Britton, & R. Ebert (San Francisco: ASP), 81
- Masci, F., et al. 2004, in ASP Conf. Ser., Vol. 347, ADASS XIV, ed. P. L. Shopbell, M. C. Britton, & R. Ebert (San Francisco: ASP), 468
- Masci, F., Makovoz, D., & Moshir, M. 2004, *PASP*, 116, 842
- Moshir, M., Fowler, J., & Henderson, D. 2003, in ASP Conf. Ser., Vol. 295, ADASS XII, ed. H. E. Payne, R. I. Jedrzejewski, & R. N. Hook (San Francisco: ASP), 181
- Moshir, M. 2002, in ASP Conf. Ser., Vol. 281, ADASS XI, ed. D. A. Bohlender, D. Durand, & T. H. Handley (San Francisco: ASP), 336
- Narron, B., et al. 2003, in ASP Conf. Ser., Vol. 295, ADASS XII, ed. H. E. Payne, R. I. Jedrzejewski, & R. N. Hook (San Francisco: ASP), 160
- Pesenson, M., et al. 2004, in ASP Conf. Ser., Vol. 347, ADASS XIV, ed. P. L. Shopbell, M. C. Britton, & R. Ebert (San Francisco: ASP), 153
- Shupe, D., et al. 2004, in ASP Conf. Ser., Vol. 347, ADASS XIV, ed. P. L. Shopbell, M. C. Britton, & R. Ebert (San Francisco: ASP), 491
- Werner, M., et al. 2004, *ApJS*, 154, 1